



MSR '24

21st INTERNATIONAL CONFERENCE ON
MINING SOFTWARE REPOSITORIES

April 15-16, Lisbon, Portugal

2024 MSR Challenge



DevGPT: Studying Developer-ChatGPT Conversations



Tao Xiao

Nara Institute of Science and Technology

Christoph Treude

University of Melbourne

Hideaki Hata

Shinshu University

Kenichi Matsumoto

Nara Institute of Science and Technology

(AI)

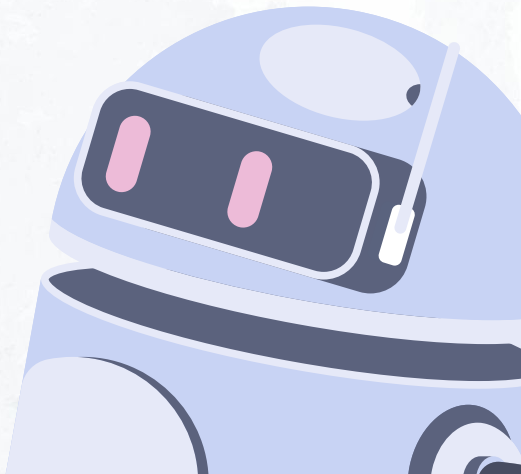


Table of contents

01 → What is DevGPT?

02 → Potential research questions

03 → Dataset structure

04 → Submission instructions

05 → Q&A

01 →

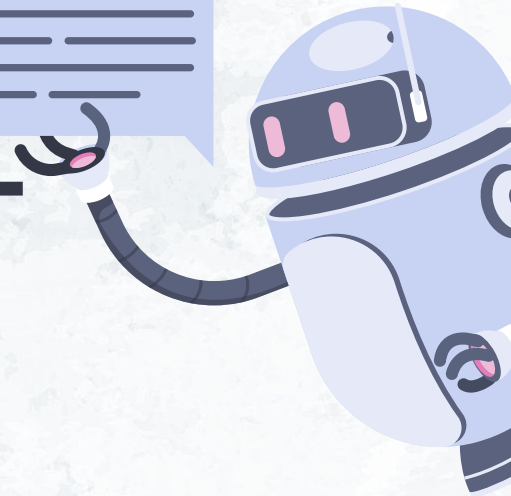
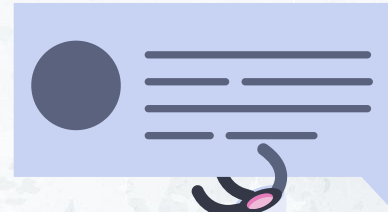
What is DevGPT?

(AI)

(DevGPT) =

Developer-ChatGPT

Conversations →



DevGPT

A curated dataset which encompasses 17,913 prompts and **ChatGPT**'s responses including 11,751 code snippets, coupled with the corresponding software development artifacts—ranging from **source code**, **commits**, **issues**, **pull requests**, to **discussions** and **Hacker News threads**—to enable the analysis of the context and implications of these developer interactions with ChatGPT.



Goal of the Challenge

Enable a comprehensive analysis of the context and implications of developer interactions with **ChatGPT**.



Example of ChatGPT Sharing

Merged FlatAHB Dump: Don't use new GUID for otherwise unchanged lines #158
Author merged 4 commits into main from dont-use-new-guid 3 weeks ago

Reviewer reviewed 3 weeks ago [View reviewed changes](#)

src/kohlrabi/unfoldedahb/unfoldedahbtable.py **Outdated**

```
299 +     if dump_equals_existing_file_except_for_guids:
300 +         ahb_line_schema = AhbLineSchema()
301 +         for line, existing_line in zip(flat_ahb.lines, existing_flat_ahb.lines):
302 +             line_copy = ahb_line_schema.load(ahb_line_schema.dump(line))
```

Reviewer 3 weeks ago · edited ▾ **Collaborator** ⋮

This is a C# pattern for deep copies isn't it?
I thought in Python there are methods to create deep copies.
asking chatGPT
ah yes there is a standard lib called `copy`
See <https://chat.openai.com/share/4ad4c1ad-6f13-4a21-a0e3-aaf75e17c556>

ChatGPT shared conversation

Reply...

besser

23e4ecf

Modified commit based on the review

Deep Copy in Python.

July 24, 2023

You are a Python expert.
How can I create a deep copy of a variable?

Certainly! In Python, if you want to create a deep copy of a variable, especially for complex objects like lists or dictionaries that contain nested objects, you should use the `'copy'` module which provides the `'deepcopy'` function.

Here's how you can use it:

```
python Copy code
import copy

original_list = [[1, 2, 3], [4, 5, 6]]
deep_copied_list = copy.deepcopy(original_list)

# Now, if you modify the nested list inside the original list...
original_list[0][0] = 99

print(original_list) # Outputs: [[99, 2, 3], [4, 5, 6]]
print(deep_copied_list) # Outputs: [[1, 2, 3], [4, 5, 6]] (remains unchanged)
```

```
30 + def _lines_are_equal_ignoring_guid(line1: AhbLine, line2: AhbLine) -> bool:
31 +     """
32 +     returns true iff the line1 and line2 are equal except for their guid
33 +     """
34 +     line1_copy = copy.deepcopy(line1)
35 +     line2_copy = copy.deepcopy(line2)
36 +     line1_copy.guid = None
37 +     line2_copy.guid = None
38 +     return line1_copy == line2_copy
```

Why DevGPT?

Understand developer usage of ChatGPT in software development.

Analyze developer questions and interaction dynamics.

Evaluate the impact on software development artifacts.

Gain insights into AI model integration in development.

Inform future AI model strategies for dev tools.



Overview of DevGPT

TABLE I
SUMMARY STATISTICS OF THE SNAPSHOT 20230831

Sources	#	Mentioned in	Shared ChatGPT Links			ChatGPT Conversations	
			# Shared Links	# Accessible Links	# Conversations with Code	# Prompts	# Code Snippets
GitHub Code File	970	Code	1386	1306	577	12320	7190
GitHub Commit	481	Message	481	477	470	1387	1569
GitHub Issue	353	Comment	264	243	160	844	739
		Description	150	138	98	854	757
		Title	3	3	3	49	77
Hacker News	195	Comment	274	240	50	837	163
		Attached URL	44	39	3	380	65
		Story	17	14	4	57	66
GitHub Pull Request	193	Description	82	81	57	494	428
		Review Thread	74	67	48	154	146
		Comment	65	60	35	351	397
GitHub Discussion	45	Comment	25	22	11	72	37
		Description	20	19	12	89	94
		Reply	6	5	3	25	23

02 →

Potential research questions

(AI)

Questions to be answered

(a) What types of issues (bugs, feature requests, theoretical questions, etc.) do developers most commonly present to ChatGPT?

(b) Can we identify patterns in the prompts developers use when interacting with ChatGPT, and do these patterns correlate with the success of issue resolution?

(c) What is the typical structure of conversations between developers and ChatGPT? How many turns does it take on average to reach a conclusion?

Questions to be answered (Cont.)

(d) In instances where developers have incorporated the code provided by ChatGPT into their projects, to what extent do they modify this code prior to use, and what are the common types of modifications made?

(e) How does the code generated by ChatGPT for a given query compare to code that could be found for the same query on the internet (e.g., on Stack Overflow)?

(f) What types of quality issues (for example, as identified by linters) are common in the code generated by Chat- GPT?

Questions to be answered (Cont.)

- (g) How accurately can we predict the length of a conversation with ChatGPT based on the initial prompt and context provided?
- (h) Can we reliably predict whether a developer's issue will be resolved based on the initial conversation with ChatGPT?
- (i) If developers were to rerun their prompts with ChatGPT now and/or with different settings, would they obtain the same results?

03 →

Dataset structure

File organizations

(a) Snapshot



9



(b) File format



`{obtained_time}_{source}_sharings.json`

(c) CSV



All shared ChatGPT links

Structure

(a) JSON



{JSON}

(b) Type attribute



[pull request, commit, hacker news,
issue, discussion, code file]

(c) General attributes



[URL, Author, RepoName,
RepoLanguage, ...]

(d) ChatgptSharing



[{URL, Status, Conversations, Mention,
HTMLContent ... } ...]

Structure (Cont.)

(e) Conversations → [{"Prompt", "Answer", "ListOfCode"}, ...]

(f) Mention → [MentionedURL, MentionedProperty,
MentionedAuthor, MentionedText, ...]

JSON files

(a) Type attributes

(b) General attributes

(c) ChatgptSharing

(e) Mention

(d) Conversations

```
{
  "Sources": [
    {
      "Type": "pull request",
      "URL": "https://github.com/validatorjs/validator.js/pull/2316",
      "Author": "Vivekkushwaha123",
      "RepoName": "validatorjs/validator.js",
      "RepoLanguage": "JavaScript",
      "Number": 2316,
      "Title": "Fixed Iranian mobile numbers code",
      "Body": "<!--\r\nAdd a descriptive title textbox above, e.g.\r\nfeat(validatorName): brief title of what has been done\r\n-->\r\n\r\nVivek Kushwaha : Fixed Iranian mobile number code\r\n\r\n<!--",
      "CreatedAt": "2023-10-07T13:10:37Z",
      "ClosedAt": null,
      "MergedAt": null,
      "UpdatedAt": "2023-10-07T13:10:37Z",
      "State": "OPEN",
      "Additions": 1,
      "Deletions": 1,
      "ChangedFiles": 1,
      "CommitsTotalCount": 1,
      "CommitSha": [
        "d798e08174f69649c6acfd2f4644e405f148127"
      ]
    },
    {
      "ChatgptSharing": [
        {
          "URL": "https://chat.openai.com/share/89797c7b-fedf-491d-8725-927820b15d88",
          "Mention": {
            "MentionedURL": "https://github.com/validatorjs/validator.js/pull/2316",
            "MentionedProperty": "body",
            "MentionedAuthor": "Vivekkushwaha123",
            "MentionedText": "<!--\r\nAdd a descriptive title textbox above, e.g.\r\nfeat(validatorName): brief title of what has been done\r\n-->\r\n\r\nVivek Kushwaha : Fixed Iranian mobile num"
          },
          "Status": 200,
          "DateOfConversation": "October 7, 2023",
          "DateOfAccess": "2023-10-12 23:25:06.598090",
          "Title": "Regex Testing Assistance",
          "NumberOfPrompts": 3,
          "TokensOfPrompts": 121,
          "TokensOfAnswers": 350,
          "Model": "Default (GPT-3.5)",
          "Conversations": [
            {
              "Prompt": "can u be my regex tester",
              "Answer": "ChatGPTCertainly, I can help you test regular expressions (regex). Just provide me with the regex pattern you want to test, and some sample text or strings that you",
              "ListOfCode": []
            },
            {
              "Prompt": "/^(\+98[\-\s]?[0]9[0-9])\d{3}[\-\s]?[0-9]{4}$/",
              "Answer": "ChatGPTIt looks like you've provided a regex pattern that appears to be designed for validating Iranian phone numbers. This regex pattern matches Iranian phone num",
              "ListOfCode": []
            }
          ],
          "HTMLContent": "<!DOCTYPE html>\n<html><head><meta charset='utf-8'/><meta content='ChatGPT: Get instant answers, find inspiration, learn something new' name='title'/><meta conte"
        }
      ]
    }
  ]
}
```


CSV file

URL	Status	MentionedURL	MentionedSource	MentionedProperty	MentionedAuthor
https://chat.openai.com/share/4bad57dd-9636-4b63-9e73-37351f51dbc9	200	https://github.com	commit	message	grnpin
https://chat.openai.com/share/1fd7ffc2-9264-45dd-943c-9fe30c182d83	200	https://github.com	commit	message	grnpin
https://chat.openai.com/share/99d2d02f-7e3b-435b-b918-9b73bd9d00c3	200	https://github.com	commit	message	grnpin
https://chat.openai.com/share/90ebe55e-bd60-47cf-b290-272980b7495e	200	https://github.com	commit	message	bh679
https://chat.openai.com/share/6b654f3c-3763-4292-92b1-3026e647e505	200	https://github.com	commit	message	4lejandrito
https://chat.openai.com/share/17a9692d-b893-47da-bca1-b46554f00962	200	https://github.com	commit	message	changchiyou
https://chat.openai.com/share/7cf2816f-13ea-404c-b0f7-ffc65583617a	200	https://github.com	commit	message	changchiyou
https://chat.openai.com/share/7cf2816f-13ea-404c-b0f7-ffc65583617a	200	https://github.com	commit	message	changchiyou
https://chat.openai.com/share/03ce2877-4d85-4e81-b70c-01d729fc3787	200	https://github.com	commit	message	RexHsu09
https://chat.openai.com/share/ce7b2d74-5552-429a-8c4f-da2cd5f908e8	200	https://github.com	commit	message	X3msnake
https://chat.openai.com/share/49eb1aff-ce9a-4373-ad28-348c2b5addcd	200	https://github.com	commit	message	Hजारadnan
https://chat.openai.com/share/58d110d6-4236-461c-b3c4-a8df6519c534	200	https://github.com	commit	message	unknowntpo
https://chat.openai.com/share/2d2dbb5d-e590-4884-a112-6de5285ccfd2	200	https://github.com	commit	message	simonw

Dataset



[snapshot_20231012](#)



[GitHub repository](#)

04 →

Submission instructions

(AI)

Submission instructions

(+) Specify which snapshot/version of the DevGPT dataset was utilized

(+) ACM Primary Article Template

(+) Latex code

```
\documentclass[sigconf,review,anonymous]{acmart}
\acmConference[MSR 2024]{MSR '24: Proceedings of the 21st
International Conference on Mining Software Repositories}{April 15–16,
2024}{Lisbon, Portugal}
```

Submission instructions (Cont.)

(+) 4 pages plus 1 additional page of references)

(+) Double-anonymous review

(+) <https://msr2024-challenge.hotcrp.com/>

(+) Cite

```
@inproceedings{ title={DevGPT: Studying Developer-ChatGPT  
Conversations}, author={Xiao, Tao and Treude, Christoph and Hata,  
Hideaki and Matsumoto, Kenichi}, year={2024}, booktitle={Proceedings of  
the International Conference on Mining Software Repositories (MSR  
2024)}, }
```


05 →

Q&A

(AI)

Thanks! →

Any questions?

Create new issues or discussions:

<https://github.com/NAIST-SE/DevGPT>



CREDITS: This presentation template was created by **Slidesgo** and includes icons by **Flaticon**, infographics & images by **Freepik** and content by **Eliana Delacour**

Please, keep this slide as attribution

